Scienstein: A Research Paper Recommender System

Bela Gipp[1], Jöran Beel[1], Christian Hentschel[2]
[1] *Otto-von-Guericke University, Dept. of Computer Science, Magdeburg, Germany*
[2] *Fraunhofer Institute for Telecommunications, Berlin, Germany*
Bela@Gipp.com, Joeran@Beel.org, christian.hentschel@hhi.fraunhofer.de

Abstract

This paper introduces Scienstein, the first hybrid research paper recommender system and a powerful alternative to currently used academic search engines. Scienstein improves the approach of the usually used keyword-based search by combining it with citation analysis, author analysis, source analysis, implicit ratings, explicit ratings and in addition, innovative and yet unused methods like the 'Distance Similarity Index' (DSI) and the 'In-text Impact Factor' (ItIF). Instead of entering just keywords, a user may provide entire documents, including reference lists as input and make implicit and explicit ratings to improve recommendations. With citation, author and source analysis, similar and related documents are easily determinable. All these techniques are managed by a user-friendly GUI.

Index Terms—DSI, Recommendation, Recommender Systems, Research paper

I. Introduction

Many scientists consider the search for related work as an extremely time-consuming part of their responsibilities. The enormity of time taken is partly caused by the increasing number of publications, which grows exponentially at a yearly rate of 3.7 % [1]. The strength of currently used academic search engines lies in finding documents containing specific keywords. Due to synonyms and unclear nomenclatures, this approach delivers in practice, often unsatisfying results.

In this paper we present Scienstein[1], a hybrid recommender system, which uses both content-based and collaborative-based techniques. We believe that this approach has the potential to alleviate the problem of finding relevant research papers. Instead of solely relying on text mining, Scienstein combines citation analysis, implicit ratings, explicit ratings, author analysis and source analysis to a recommender system with a user-friendly GUI. Currently, Scienstein is in the development stage and open for cooperation.

The first part of this paper gives an overview of related work including a discussion of the advantages and disadvantages of existing approaches. The main part introduces Scienstein and discusses the technologies used. The focus lies on a hybrid recommender approach, which combines content-based and collaborative-based approaches. It shows that many of the disadvantages of existing systems become obsolete by combining known concepts with new ones. The last part of the paper gives insights into the usage of the software by illustrating its functionality with screenshots.

[1] www.scienstein.org

II. Related Work

In practice, research paper recommender systems do not exist. However, concepts have been published and partly implemented that could be used for their realisation. Some authors suggest using collaborative filtering and ratings. Ratings could be directly obtained by considering citations as ratings [2] or implicitly generated by monitoring readers' actions such as bookmarking or downloading a paper [3], [4]. Citation databases such as CiteSeer apply citation analysis (e.g. bibliographic coupling [5] or co-citation analysis [6], [7]), in order to identify papers that are similar to an input paper [8]. Scholarly search engines such as Google Scholar focus on classic text mining and citation counts.

Each concept does have disadvantages, which limits its suitability for generating recommendations.

For example, citation analysis cannot identify homographs[2], and not all research papers are listed in citation databases. Likewise, reference lists can contain irrelevant entries caused by the Matthew Effect[3], self citations[4], citation circles[5] and ceremonial citations[6].

Other problems pop up with text-based analysis, which has to cope with unclear nomenclatures, synonyms or context depending on the meanings of words. Accordingly, text-based

[2] Homographs describe authors with identical names. As a result, citation analysis sometimes cannot assign a research paper to its correct author [9].

[3] The Matthew Effect describes the fact that frequently cited publications are more likely to be cited just because the author believes that well-known papers should be included [10].

[4] Sometimes self citations are made to promote other publications of the author, although they are irrelevant for the citing publication [11].

[5] Citation circles occur if citations were made to promote the work of others, although they are pointless [12].

[6] Ceremonial citations are citations that were used although the author did not read the cited publication [9].

recommender systems cannot identify related papers if different terms are used.

Collaborative filtering in the domain of research paper recommendation is criticised for various reasons. Some authors claim that collaborative filtering would be ineffective in domains where more items than users exist [13]. Others believe that users would be unwilling to spend time for explicitly rating research papers [2]. Problematic with implicit ratings is that for obtaining the required data, continuous monitoring of the researcher's work is necessary, which raises privacy issues[7]. In general, collaborative filtering has to cope with the possibility of manipulation. Another drawback is that a critical mass of ratings and users is required to receive useful recommendations.

III. SCIENSTEIN: A HYBRID RECOMMENDER SYSTEM

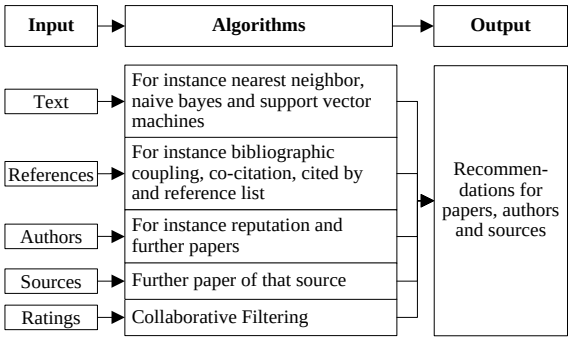

Figure 1: Scienstein's approach to recommend research papers

Since all current search engines and concepts for research paper recommender systems focus mainly on one approach (text analysis, citation analysis or ratings), each concept suffers the disadvantages mentioned above. The Scienstein project aims to combine the already known concepts with new ones in order to create a holistic research paper recommender system. By combining different concepts, many disadvantages become obsolete. Scienstein's approach to recommend research papers is illustrated in Figure 1. With Scienstein, users may provide one or several of the six inputs (text, references, authors, sources, ratings or documents), adjust the algorithms to their needs[8], and receive recommendations for research papers. Further plans for the future include broadening Scienstein's functionality so that authors, journals or conferences can also be recommended.

In addition to the technical side, Scienstein offers a user-friendly GUI so that the complex technical possibilities can be handled without expert knowledge in formulating search queries etc.

[7] If document usage is permanently monitored, employers with access to the usage data could, for instance, draw conclusions about the researchers' working times and productivity.

[8] For instance, put more weight on finding papers similar to the input document or finding papers published by the same or a similar author.

IV. SCIENSTEIN'S CITATION ANALYSIS

Scienstein combines four approaches of citation analysis to identify papers that are similar to a given input paper (see Figure 2 for illustration). The 'cited by' approach considers papers relevant that cite the input document (see Figure 2, documents A and B). The 'reference list' approach considers papers relevant that were referenced in the input document (see Figure 2, documents C and D). 'Bibliographic coupling' considers papers relevant that cite the same article(s) as the input document (see Figure 2, document BibCo). With 'co-citation analysis', papers are considered relevant that were cited by those papers that were also cited by the input document (see Figure 2, document CoCit).

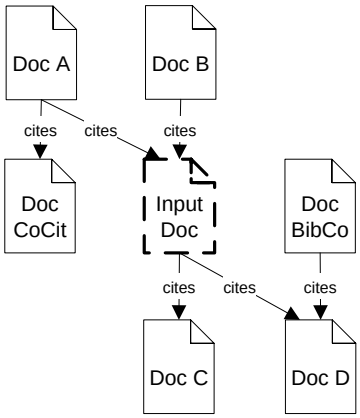

Figure 2: Co-citation, bibliographic coupling, cited by and reference list

To rank results, Scienstein applies what we call 'in-text citation frequency analysis' (ICFA) and 'in-text citation distance analysis' (ICDA).

ICFA analyses the frequency with which a research paper is cited within the citing document. We developed the 'In-text Impact Factor' (ItIF), which represents the number of citations referring to a certain document divided by the overall number of citations (see Figure 3 for illustration). The sum of all ItIF of one document always adds up to 1. The higher the ItIF, the closer related are the input document to the cited document.

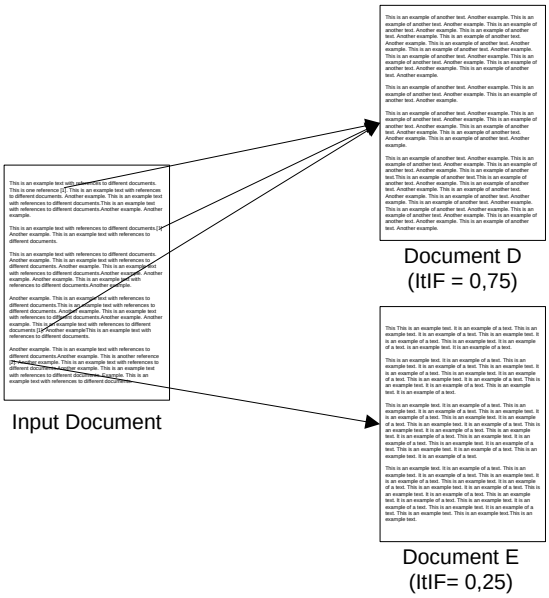

Figure 3: In-text Impact Index (ItIF)

ICDA analyses the distance between references within a text to determine the degree of their similarity (see Figure 5 for illustration). The idea is that the more similar two documents are, the more likely they are closely referenced in other research papers. For Scienstein, the 'Distance Similarity Index' (DSI) was developed, which calculates the similarity of two documents based on the citation distance. If two references occur in the same sentence, the referenced documents are likely to be very similar and the DSI is 1. If they occur in the same paragraph the DSI is ½. The other values used are shown in Table 1.

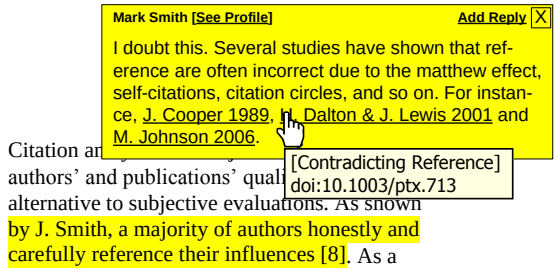

Figure 4: Collaborative Annotations and Links

First experiments with in-text frequency and distance analysis delivered promising results. However, further research needs to be performed for optimizing the algorithms and for identifying the right weighting of variables, which seem to depend on the publication's research field.

Table 1: Distance Values

Occurrence	Value
Sentence	1
Paragraph	1/2
Section	1/4

Occurrence	Value
Chapter	1/8
Other	1/16

In addition to classic references, Scienstein analyses references that were added by users and that we call 'collaborative links' [14]. These links may, for instance, occur in collaborative annotations and can be classified as contradiction, correction, supporting, or addition/improvement (see Figure 4). In contrast to references, the links may point to publications that were published after the paper or were unknown to the author and hence provide valuable information to the readers. For recommendation purposes, the links' classifications are important. With citation analysis based on classic references it can only be determined that two documents are related somehow. With classified collaborative links it can be expressed how they are related.

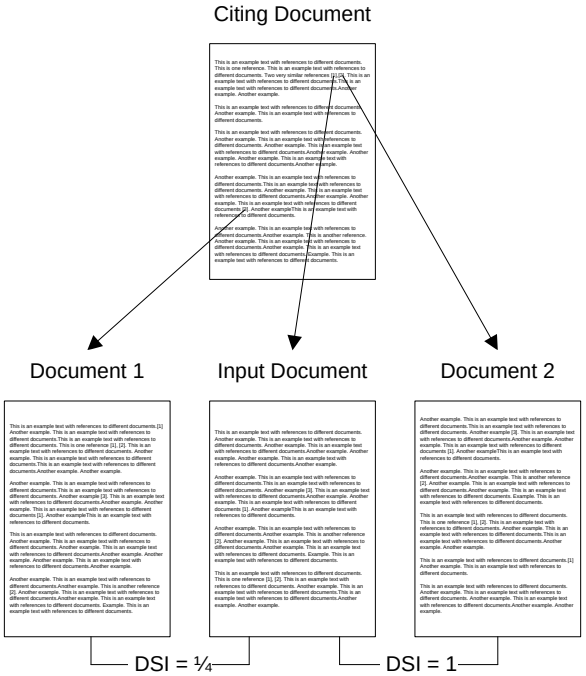

Figure 5: Distance Similarity Index

V. Scienstein's Author and Source Analysis

To find further potentially relevant papers, Scienstein uses a simple, but nevertheless in practice unapplied method. Those papers are considered relevant that were published by the same author or source (e.g. journal) as the input document. The basic principle is illustrated in Figure 6. Additionally, author and source analysis can be used to rank recommendations, for instance, by reputation. In Scienstein, the user can decide the way reputation is measured. Besides common standards such as the impact factor or h-index, users can define the way reputation is measured themselves, e.g. by implicit and explicit ratings or combinations.

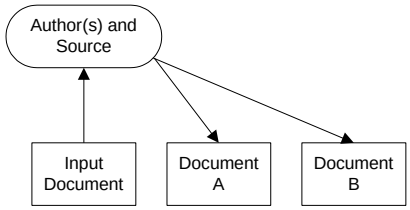

Figure 6: Author and Source Based Analysis

VI. Scienstein's Text Mining

With regard to text mining, Scienstein basically uses existing techniques and only offers some additional features such as the possibility of classifying papers according to details given in the acknowledgements. This way, for instance, research projects supported by lobby groups can be easily identified, if mentioned in the acknowledgements.

Additionally, Scienstein considers data gathered by collaborative annotations and classifications [14, 15]. Collaborative annotations are in-text comments made by the readers (see Figure 4). Collaborative classifications are similar to tags, but more structured. In the current prototype of Scienstein, users can add tags in three main categories: field of research, research methods and research details. In case of interdisciplinary work for each category, several tags can be assigned. Additionally, further categories can be created. For instance, it might be useful to classify publications about archaeological sites according to their geographic location. This would allow the development of, for instance, a Google Maps extension so that the user can zoom into sites to get relevant publications listed.

The advantage of collaborative annotations and classifications is that new terms can retrospectively be associated to documents. For instance, Goldberg et al. published in 1992 the idea of what we call today a recommender system [16]. However, the term 'recommender system' was actually coined two years later by Resnick et al. [17].

VII. Scienstein's Document Rating

As explained, some authors consider collaborative filtering and explicit ratings as unsuitable for recommending research papers. However, we do not know of any studies supporting their assumptions. In contrast, we believe that for the majority of users, the costs of participating would be lower than the benefits for the following reasons [18].

- Explicit ratings improve a user's own recommendations accuracy
- Explicit ratings deliver document management functionality by serving as extended memory for a user's preferences
- Explicit ratings pleases a user by allowing him/her! to contribute to an advancing community
- Explicit ratings provide the satisfaction of having one's own opinion voiced and valued

Even if only very few users participate in the explicit rating of research papers, we believe that these ratings still deliver valuable information complementing the other approaches.

Table 2: Actions monitored in Document Usage Mining

View Document Details	Edit Document Details
Read Abstract	Highlight passages in PDF
Bookmark Document	Create Bookmark within PDF
View Coll. Annotations	Add Coll. Annotations
View Coll. Ratings	Add Coll. Ratings
View Coll. Links	Add Coll. Links
View Coll. Classifications	Add Coll. Classifications
View Bibliography	Send/Recommend to friend
Download	Print
Read	Follow Recommendations
View Related Documents	Reference Document

In addition to explicit ratings, Scienstein generates implicit ratings by monitoring 22 user actions (see Table 2). We call the process of monitoring the user's actions on a document 'document usage mining'. The underlying assumption is that intensively studied documents or paragraphs in documents are more valuable for the user than documents that were, for example, closed after a few seconds.

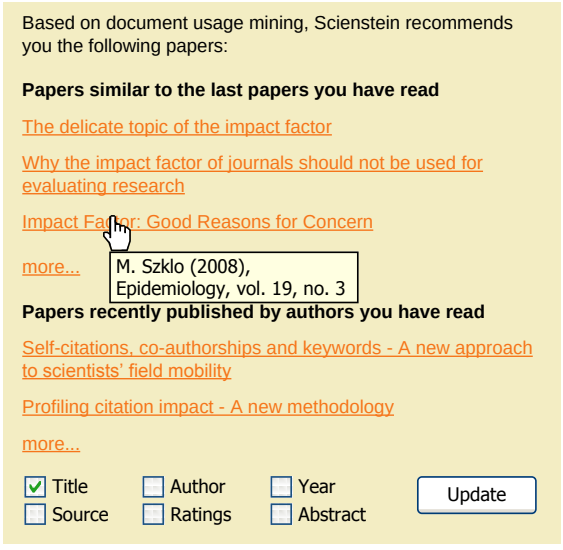

Figure 7: Document usage mining based recommendations

For generating implicit ratings and recommendations (see Figure 7), Scienstein weighs each activity based on the user's past behaviour. For instance, some users print every potential relevant document, whereas others only print documents after careful inspection. In the latter case, the activity 'print' would be assigned a higher weight.

VIII. User Interface

To support the user in managing the information flood resulting from the various technical possibilities, Scienstein set store by the development of a user-friendly GUI. A selection of important concepts is presented in the following.

A. *Cockpit view*

The 'Cockpit View' is the core of the Scienstein recommender GUI (see Figure 9). It consists of the graphical representation of recommended documents, various controls to filter them and a context dependent legend.

The graphical view shows the recommended documents, whereas the size of the displayed documents depends on the degree they fulfill the settings made on the right. Recommended documents are grouped according to their classification. The classification is based on the journal, keyword analysis, reference analysis and tags assigned by users. By moving the mouse over a document, further information is displayed in a yellow box. Besides obtaining a summary of relevant information the user has the possibility to rate the document or write annotations. If a document is positively rated by clicking on the green check mark it will be marked as relevant and similar documents are recommended by being immediately enlarged.

B. *Relevance Selection*

Three possibilities exist in Scienstein to build up 'search queries'. First, an ordinary keyword-based search using Boolean operators can be performed. Secondly, a research paper can be uploaded to perform a keyword and reference analysis. The third possibility is to receive recommendations by entering arbitrary text and references into a text box and marking relevant and irrelevant content by a red or green virtual highlighter (see Figure 9).

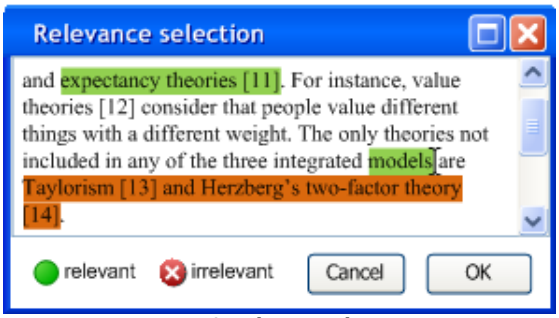

Figure 9: Relevance selection

If desired, the approaches can be combined for further filtering. The screenshot below illustrates this method. The green-marked keywords and references are then included and the red-marked content is excluded from recommendations.

C. *Project Selection*

Researchers often work on different projects at the same time and hence need recommendations in different fields. In the case of explicit and implicit ratings, a recommender system needs to consider under which circumstances respectively during which project a rating was performed.

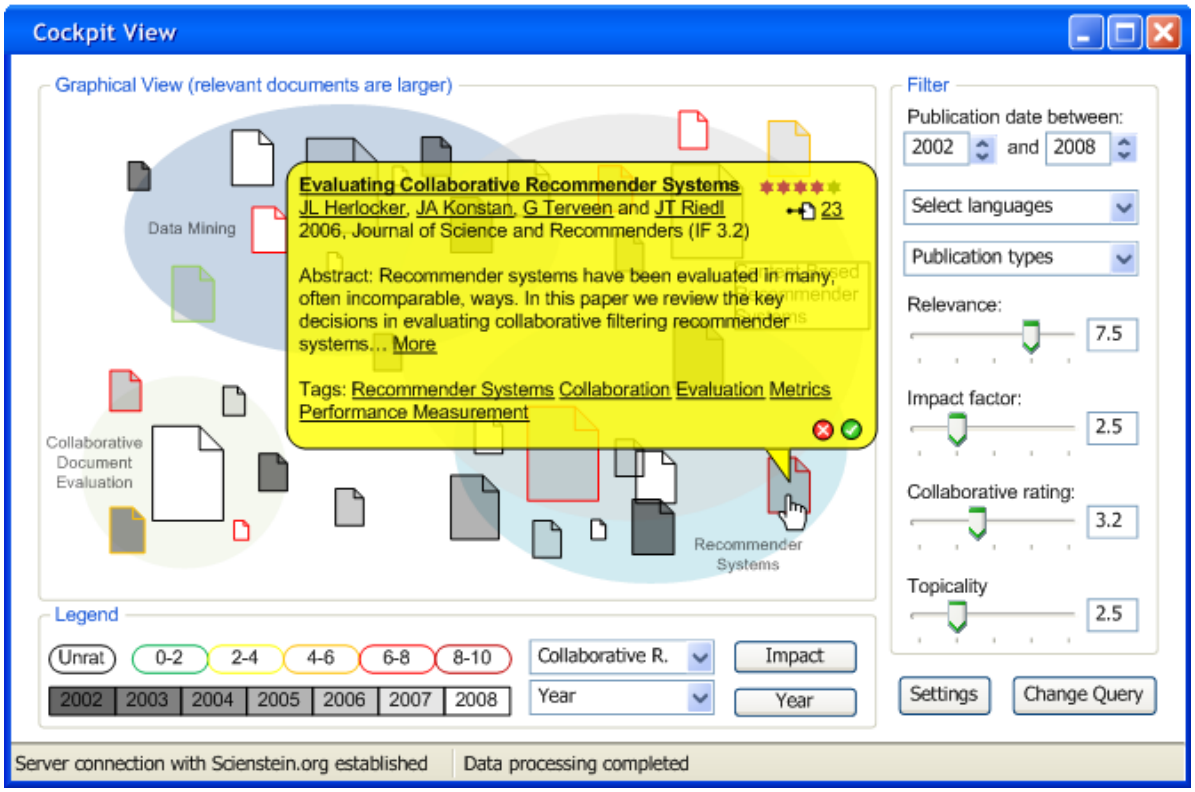

Figure 8: Cockpit View

Therefore, different projects can be defined and new ones can be derived from existing ones. These profiles can be published to assist other researchers in finding relevant literature (see Figure 10).

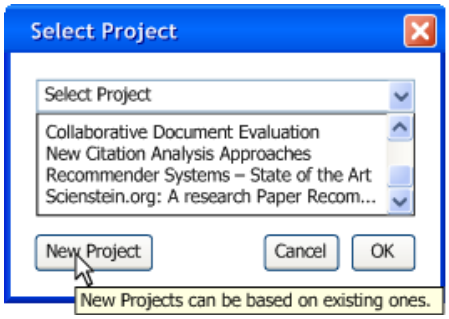

Figure 10: Project Selection

IX. Conclusion

In this paper, Scienstein, the first hybrid recommender system for research papers, was introduced. Scienstein aims to be a powerful alternative to academic search engines by not solely relying on keyword analysis, but by additionally using citation analysis, explicit ratings, implicit ratings, author analysis, and source analysis. Although some of the utilized methods have been known for decades, they have not been applied in the context of research paper recommender systems. Other approaches such as the 'in-text distance similarity index' or collaborative annotations, classifications and links were developed exclusively for Scienstein. The combination of all approaches is critical since each approach possesses disadvantages that can only be overcome by combining them.

However, many questions remain unanswered, for instance regarding non-technical aspects like privacy concerns resulting from implicit and explicit ratings. Further research in this field will be performed by the Scienstein team, which welcomes other researchers to participate.

X. References

[1] May, R. M. 1997. The Scientific Wealth of Nations, Science, vol. 275, no. 5301, pp. 793-796.

[2] Torres, R. McNee, S. M. Abel, M. Konstan, J. A.. and Riedl, J. 2004. Enhancing Digital Libraries with TechLens, Proceedings of JCDL'04, pp. 228-236.

[3] Pennock, D. M. Horvitz, E. Lawrence, S. and Giles, L. C. 2000. Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach, in Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (San Francisco).

[4] Middleton, S.E. Shadbolt, N. R. and De Roure, D. C. 2004. Ontological User Profiling in Recommender Systems, ACM Transactions on Information Systems (TOIS), vol. 22, no.1, pp. 54-88.

[5] Fano, R. M. 1956. Information theory and the retrieval of recorded information, in Documentation in Action, Shera, J. H. Kent, A. Perry, J. W. (Edts), New York: Reinhold Publ. Co., pp. 238–244.

[6] Marshakova, I. V. 1973. System of document connections based on references, Nauchno-Tekhnicheskaya Informatsiya, vol. 2, no. 6, pp. 3–8.

[7] Small, H. 1973. Co-citation in the scientific literature: a new measure of the relationship between two documents, Journal of the American Society for Information Science, vol. 24, pp. 265–269.

[8] Giles, C. L. Bollacker, K. D. And Lawrence, S. 1998. CiteSeer: an automatic citation indexing system, In Digital Libraries 98 - The Third ACM Conference on Digital Libraries, pp. 89-98.

[9] Meho, L. I. 2006. The Rise and Rise of Citation Analysis, Physics World, http://arxiv.org/abs/physics/0701012.

[10] Merton, R. K 1968. The Matthew Effect in Science, Science, vol. 159, no. 3810, pp. 56-63.

[11] Tagliacozzo, R. 1977. Self-citations in scientific literature, Journal of Documentation, vol. 33, no. 4, pp. 251-265.

[12] Garfield, E. and Welljams-Dorof, A. 1992. Citation data: their use as quantitative indicators for science and technology evaluation and policy-making, Science & Public Policy, vol. 19, no. 5, pp. 321-327.

[13] Agarwal, N. Haque, E. Liu, H. and Parsons, L. 2005. Research Paper Recommender Systems: A Subspace Clustering Approach, In Advances in Web-Age Information Management, Springer: Heidelberg.

[14] Beel, J. and Gipp, B. 2008. Collaborative Document Evaluation: An Alternative Approach to Classic Peer Review. In Proceedings of World Academy of Science, Engineering and Technology, vol. 31, pp. 410-413.

[15] Beel, J. and Gipp, B. 2008. The Potential of Collaborative Document Evaluation for Science, the 11th International Conference on Digital Asian Libraries (ICADL 2008), published in G. Buchanan, M. Masoodian & S. Cunningham (Eds.), Digital Libraries: Universal and Ubiquitous Access to Information of Lecture Notes in Computer Science, vol. 5362, DOI 10.1007/978-3-540-89533-6, ISSN 0302-9743, pp. 375-378, Springer-Verlag Berlin, Heidelberg.

[16] Goldberg, D. Nichols, D. Oki, B. M. and Terry, D. 1992. Using Collaborative Filtering to Weave an Information Tapestry, Communications of the ACM, vol. 35, no. 12, pp. 61-70.

[17] Resnick, P. Iacovou, N. Suchak, M. Bergstrom, P. and Riedl, J. 1994. GroupLens: An open architecture for collaborative filtering of Netnews, in Proc. ACM conference on Computer supported cooperative work, Chapel Hill, North Carolina, United States ACM Press.

[18] Harper, F. Li, X. Chen, Y. and Konstan, J. 2005. An Economic Model Of User Rating In An Online Recommender System, in Proceedings of the 10th International Conference on User Modeling, Edinburgh, UK.